**Response to reviews:**

1. Are the objectives and the rationale of the study clearly stated?

Reviewer #1: The objectives and rationale for this study are clearly described.

Reviewer #2: Yes.

1. If applicable, is the application/theory/method/study reported in sufficient detail to allow for its replicability and/or reproducibility?

Reviewer #1:

Please provide additional detail in the methods section on how the age-length keys (ALKs) were constructed. Specifically, how was the bin structure selected (single or multi-centimeter bins)? Was this choice adjusted to be consistent with the range of lengths observed for each species and/or the specific bin structure that is used in the stock assessments?

**Response: In the revision we have included additional text in an effort to clarify our expansion methods and to provide additional detail on how the ALKs were constructed. Specifically, we have indicated in the methods section that we are binning the length frequency (and subsequent ALK) to 1 cm bins, that the bins were across the range of lengths for each species, and followed the bin structure used in the assessments. We also note in the methods that the framework we have constructed allows for any bin structure. To further explore the potential effect of bin structure on our results, and to generalize our results, we have included alternative length bins in the analysis in what we term the ‘Length bin treatment’. In this treatment we also include 2 cm and 5 cm length bins in addition to the 1 cm length bin base case. We show, that while the magnitude of age composition ISS does increase slightly as the bin size is increased, the general results of decreasing age composition ISS as ageing error and growth variability were consistent across all the bin sizes.**

It is somewhat unclear from the methods, but I think that the “total” or combined-sex results were created by summing the male and female age compositions. Would the ‘total’ age composition be better estimated by combining males and females prior to creating the age-length key and therefore not including the bootstrapped variability of each sex separately in the total? Would this be the appropriate method if the stock assessment were intending to use the aggregate distribution? With a larger sample size, albeit possibly also with more growth variability, would the input sample sizes would be larger using this method?

**Response: We agree with the reviewer that we left it unclear as to how the combined-sex results were created and have added text to the revised methods section that describe that the combined-sex results were created by summing the male, female, and unsexed expanded population numbers-at-length and -age to then compute the total length and age compositions, as the reviewer noted. The primary reason this is done is to account for the potential differences in growth between sexes (for example, flatfish species). Albeit, the reviewer makes an interesting point that the increased sample size could translate to an increased ISS if the data were combined prior to expansion. We were curious about this and ran an example for a two species that don’t exhibit difference in growth between sexes (Gulf of Alaska Pacific ocean perch and walleye pollock). We found that the age composition ISS did indeed increase slightly with data that is combined across sexes prior to length and age expansion, however, the main results of age composition ISS decreasing as additional sources of uncertainty are included was consistent whether creating the total composition before or after length and age expansion. This was an interesting enough result that we have included this as additional analysis in what we term the ‘Aggregration treatment’ in the revised manuscript.**

Were the age samples in this study collected randomly in each year or were there stratified targets by length? If length-stratified in some years (which would seem best for the ALK approach), but random in other years (perhaps best for building age compositions without an ALK and having more variability when used with an ALK) would the interaction with ageing imprecision be different? If so, should the results be delineated between years with different sampling methods?

**Response: The reviewer is correct that sampling methodology has changed across time as it pertains to age collections on AFSC bottom trawl surveys. It has been the case that for species assessed with SCAA models at AFSC the transition has been from length-stratified to random sampling, where currently all species assessed with SCAA models at AFSC are collected randomly. However, we note that the collection of otoliths isn’t completely random, at least in the ‘simple random sampling’ sense. This is because the number of otoliths within any given haul are capped at some level for all these species (for example, 2 fish per haul for arrowtooth founder, as compared to 5 fish per haul for Pacific ocean perch) and not collected proportionally to catch. There are two primary complications that this creates when considering building an age composition without an ALK. First, these haul level samples would need to be weighted in some form, whether by catch or cpue of the haul. Second, at the strata level the number of otoliths collected is much smaller than the length frequency collections, so expansion by some sort of catch weighting at the strata scale would be problematic when only using age data. This remains true at any spatial scale of age versus length frequency sampling. To overcome these complications would involve developing new expansion methods, which we have not done in the revision as we feel whatever expansion methods we would develop would not mimic what is done at AFSC, and may not mimic any methods employed anywhere else. We would also suggest, that regardless of the age sampling methodology employed (length-stratified or random) that one would consider using an ALK expansion approach. Given the large difference in the magnitude of sampling between length frequency and age on most surveys one should consider leveraging the information contained in the length frequency when developing age composition through the ALK approach. To the reviewers main point, of whether ageing imprecision would have a different impact on ISS if age composition were computed through an ALK or not, we hypothesize that the impact would be similar, at least in the sense that the ISS would decrease with the added variability that ageing imprecision adds. However, we have not tested that here explicitly due to the above complications.**

1. If applicable, are statistical analyses, controls, sampling mechanism, and statistical reporting (e.g., P-values, CIs, effect sizes) appropriate and well described?

Reviewer #1:

The bootstrap approach is appropriate and the manuscript extends previous work on input sample size calculations from length and age compositions for use in fisheries stock assessment. My biggest concern is with regard to the pooling across years of the age-length key and the influence that may have on the results: A primary focus of the manuscript is the proportional contribution of growth variability to the reduction in input sample size. However, the method uses a ‘global’ ALK pooled across all years and areas. I’m assuming that assessments would normally use a year-specific key, which would not have the effects of interannual growth differences and/or strong cohorts in a single key pooled over time. If the pooled approach does not match what is done in assessments, it would seem to overestimate the reduction in input sample size due to growth. Would it be possible to add an example to the analysis comparing the results based on pooled vs annual ALKs for one of the species with a large growth effect on the input sample size (e.g. a gadid)?

**Response: The reviewer makes a very important point here, that the impact of growth variability could be over-estimated when using pooled growth data. To investigate this we included an additional analysis in the revision that we term the ‘Growth data treatment’ that evaluates pooling growth data across time or using annual data when implementing growth variability. While the review suggested doing this analysis for an example gadid species, we ran the analysis on all the stocks considered so that comparison could be made across species types. We show (as the review hypothesized) that the impact of growth variability was sensitive to how the growth data was pooled (whether using annual data or pooling the data across survey years), particularly for gadids.**

Reviewer #2:

Additional details about the estimated relationships between nominal and input sample sizes should be added as noted in the review.

**Response: As further discussed below, in the revision we have presented the relationship between age composition ISS per age sample with nominal sample size in addition to the relationship with the number of sampled hauls, including the statistics for linear fits and note that we have not included a table of average ISS per age sample or sampled haul due to the poor relationship between the two quantities.**

1. Could the manuscript benefit from additional tables or figures, or from improving or removing (some of the) existing ones?

Reviewer #1:

I think it would be helpful to include a figure/panels showing results similar to the upper panels of figure 5, but comparing with the age range in the observed data for each species and also the number of length bins with data for the ALK for that species. Perhaps these have more of a clear effect than the CVs?

**Response: In the revision we have changed this figure (now Figure 9) following the reviewers suggestion. We now show the relationship between relative ISS and the age range (as an indication of longevity) and length range (as an indication of growth). We appreciate the reviewers suggestion, because these results do show a consistent relationship across the species types where the CVs really didn’t help explain any of the results we found.**

Figure 2 (and other boxplots): What is shown by these boxplots - median, 25th, 75th and range?

**Response: Yes, these boxplots show the median, 25th and 75th percentiles, and the whiskers shown 1.5 x the inter-quartile range. In the revision we have included description of what is shown in the boxplots to the figure captions.**

Figure 4, upper panel: These axes should not extend below, as that space has no meaning. Also, the ellipses need a description in the caption: why do they extend far below the points (are they forced to cover 0,0)?, do/should they have 95% coverage of the points?

**Response: In the revision, rather than use ellipses (which we used in the original version to illustrate trends and overlap among the species types) we present linear models fit to these data. This avoids the issue of the ellipse extended into negative numbers and interpretation difficulties of what they are showing while still illustrating the trends we discuss.**

Figure 5 lower panel: These results seem inconsistent with figure 3 as the whiskers all extend above a value of 1.0 (but not in Figure 3). Is it possible that the bootstrapped values and not the annual harmonic means are being plotted here?

**Response: In the original version of these figures we presented the ISS after it was averaged across survey years for each stock in Figure 3 while showing the annual ISS values in Figure 5 (and we failed to highlight this difference in the original figure captions). In the revised version we have decided to present the annual ISS values in all the figures, and we note that in some instances the inclusion of ageing error and growth variability can result in a larger value of ISS, although this happens in a small number of cases and the general result is a decrease in ISS as these sources of uncertainty are included.**

Reviewer #2:

Yes, it would be useful to provide a table with summary statistics for the results shown in Figure 4.

**Response: In the revision we have included (1) linear relationships with *R*2 values that indicate the strength of relationships, (2) a figure that compares the age composition ISS with nominal sample size, and (3) statistics in text that report the range and variability in the mean age composition ISS per haul or age sample. We discuss below (and in the revision) that these relationships are weak and variable and we recommend that assessment scientists perform the bootstrap procedure rather than scale the number of hauls sampled or the nominal sample size to set ISS in their SCAA models. To that end, we have not included a table like this in the revision as we feel it could be misused to set ISS.**

1. If applicable, are the interpretation of results and study conclusions supported by the data?

Please provide suggestions (if needed) to the author(s) on how to improve, tone down, or expand the study interpretations/conclusions. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Yes. But see comments on the use of the pooled ALK.

**Response: As described above, we have included additional analysis in the revision that investigates the pooling of growth data when growth variability in included.**

Reviewer #2: Yes.

1. Have the authors clearly emphasized the strengths of their study/theory/methods/argument?

Reviewer #1: Yes

Reviewer #2: Yes

1. Have the authors clearly stated the limitations of their study/theory/methods/argument?

Reviewer #1:

Per the above comments on the ALK approach, the relative effect of growth on input sample size may be overestimated; however an example comparison for one species with large variability due to growth could address this.

**Response: As described above, we have included additional analysis in the revision that investigates the pooling of growth data when growth variability in included. We thank the review for this suggestion, as we feel it has helped to make this manuscript stronger.**

Reviewer #2:

No, they need to provide more context about alternative approaches to accounting for ageing error and growth variability in stock assessments.

**Response: In a response below we address how we incorporated additional analysis to present the results as they pertain to CAAL data. We also provide discussion below and in the revision to support our recommendation that ageing error and growth variability be taken into account when using ISS to weight either age composition or CAAL in SCAA models. We appreciate the reviewer’s suggestion on this particular topic, as we feel that including CAAL results help to generalize our methods and results presented in this study.**

1. Does the manuscript structure, flow or writing need improving (e.g., the addition of subheadings, shortening of text, reorganization of sections, or moving details from one section to another)?

Reviewer #1: The manuscript is well written and organized.

Reviewer #2: The flow and structure are fine.

1. Could the manuscript benefit from language editing?

Reviewer #1: No

Reviewer #2: No

Reviewer #1: Minor editorial corrections:

Lines 153-155: Which years were included in this analysis?

**Response: in the revised Data section we have included the years for each survey that were included in the analysis.**

Line 339: “Bias”

**Response: in the revision ‘bias’ has been replaced with ‘biased’.**

Line 380: Instead of “half-way”, perhaps consider “partially” since the relative magnitude of the effects from ageing error in the assessment and in the ISS bootstrap have not been quantified.

**Response: we have replaced ‘half-way’ with ‘partially’ in the revised manuscript.**

Line 400: “as example”

**Response: we have removed ‘as example’ from the revision.**

Line 403: “has largest” “and smallest”

**Response: in the revision we have replaced ‘largest’ with ‘a greater’ and ‘smallest’ with ‘less impact’.**

Reviewer #2: Big picture comments

The manuscript provides a valuable estimate of adjustments that should be made to account for input sample sizes (ISS) in statistical catch-at-age models in which length compositions are converted to age compositions using an age-length key. This approach is used at the Alaska Fisheries Science center where the authors work and in many other places. The manuscript is technically sound and surely valuable for improving the assessments at AFSC. However, it has some major gaps when it comes to providing context for this approach.

First, I believe the manuscript does not adequately compare their approach to the alternative in which length compositions are input directly into the model and fit to expected compositions calculated from a growth curve and estimated variability in length at age each. That alternative approach allows ageing error (as well as potentially process error in growth and selectivity) to be incorporated directly into the model, and was used by Thorson and Haltuch (2019) as well as in the assessments of all the stocks explored by Stewart and Hamel (2014). The paragraph starting on line 368 suggests that incorporation of ageing error in the calculation of expected age compositions addresses only half of the problem, but when length compositions are used directly, there is no need for incorporation of ageing error or growth variability in the calculation of input sample size.

**Response: The reviewer makes an important point here, that not all assessments use expanded marginal age composition data in the assessment model, but may use the expanded length composition data in conjunction with CAAL data so that the assessment can estimate growth and expand the age composition data internally. Based on this review, we developed functions in order to provide bootstrap estimates of ISS for CAAL data (and to our understanding, this is the first manuscript to present CAAL bootstrapped ISS in the literature). We show that the main results of decreasing ISS when ageing error and growth variability are included also holds for CAAL ISS. We will note that in the use of CAAL data in an assessment, an ISS is still required in order to fit this data, even though the age composition data has been internally expanded (where ageing error and growth can be taken into account). If CAAL data is being fit in a model we suggest that the same reasoning that we provide in this paragraph for externally expanded age composition data would also hold true when using CAAL data, that when fitting the estimated age composition proportions from a model (which would include ageing error and in some cases growth variability) to the observed age composition proportions (which include ageing error and growth variability based on the way they are collected and expanded) the uncertainty that is used to fit the model to the data should also take ageing error and growth variability into account. This would hold true when using CAAL data because this data also inherently includes ageing error and growth variability. However, if the model does not fit age composition or CAAL data but only fits length composition data, then yes, the reviewer is correct that there is no need for age composition ISS, as age composition wouldn’t be a part of the model’s likelihood components (using ‘likelihood’ to refer to the negative log-likelihood here). We have included discussion in the revised manuscript that addresses use of CAAL data in a model based on this review.**

Second, the manuscript does not address other data weighting approaches, such as the iterative approach developed by Francis (2011) or those that include estimated parameters promoted by Thorson (numerous, including 2023 paper co-authored by the lead author of this manuscript). It would be useful to see how the down-weighting calculated via a model-based data-weighting method compared to the adjustments proposed in the manuscript.

**Response: The reviewer is correct that we have not investigated the consequences of reductions in age composition ISS to SCAA model predictions, whether using a ‘self-weighting’ procedure or not. Our focus in this manuscript is to present the consequences of additional sources of error on ISS as an intermediate step to performing this exact investigation. In the original version of the discussion we included text that discussed the possible implications of mis-specifying ISS in SCAA models, in the revised version we also include text that discusses possible implications of the ‘self-weighting’ methods, specifically that the starting values for these weighting methods matters (that while in theory one may possibly expect the self-weighting methods to correct for misspecification, but that in practical implementation of these methods the starting values for ISS can have consequences on the results). We appreciate this review, and note that we plan to continue with this line of investigation, specifically to study with simulation whether bootstrap ISS is a useful statistic to use for weighting of composition data as compared to ad-hoc methods as well as model-based methods.**

Minor issues

Line 54: “obtained from hauls within a trip that is targeting a specific species” Perhaps this is the norm in Alaska, but in much of the world it’s more common to sample trips, not hauls, and most fisheries do not target a single species.

**Response: In the revision we have adjusted text to reflect that either hauls or trips may be sampled and that multiple species may be targeted.**

Line 66: “The statistical weight assigned to annual composition data…” The statistical weight is often a combination of one of the methods listed and some adjustment applied within the model according to some data-weighting scheme.

**Response: When considering data weighting of all the various sources of data integrated into an SCAA model the reviewer is correct, that some adjustments may be made within the model. However, in this particular section we are specifically referring to the ISS used to weight the composition data. To address this comment we have removed ‘statistical’ in the revision.**

Line 81: the proposed naming convention is valuable and welcome

**Response: We thank the reviewer and agree, in the literature we have gotten to a point that a number of terms can be used to describe the same aspect when speaking about composition data.**

Line 97: “duet o” to “due to”

**Response: we have made this correction in the revision.**

Line 108: “a number of studies” is supported only by a single reference: “Liao et al. (2013).

**Response: In the revision we have removed ‘a number of’ and have added ‘e.g.,’ prior to the reference.**

Line 212: “Growth variability was incorporated by resampling from all lengths associated with a given age and sex.” This is confounded with ageing error. Even if there were zero variability in growth, this calculation would show a range of lengths associated with a given age due to ageing error except for ages beyond the point where growth had reached an asymptote.

**Response: The reviewer makes a really good point here. However, we will note that this confounding is also included in the observations that we make (which can’t be disentangled) and that this would be an additional source of uncertainty that should be taken into account when using these data in SCAA models. This was a primary reason for why we included uncertainty scenarios that only included ageing error or growth variability so that we could investigate these sources of uncertainty independently from each other as well as understand the impacts of including both at the same time.**

Line 303: “A positive relationship is observed, by species group, between the number of age samples taken per haul and the age composition ISS per sampled haul (top panel of Figure 4).” It would be useful to provide a table with summary statistics for the results shown in Figure 4: the average input sample size per nominal age sample and/or the average ISS per haul. This would make it easier to compare to the results of Stewart and Hamel (2014) and to adjust nominal sample sizes without repeating the bootstrap procedure.

**Response: In the revision we have included relationships between both the number of sampled hauls and nominal sample size. In particular, we fit linear relationships that relate ISS to either the nominal sample size or the number of sampled hauls, with *R*2 values to give an indication of the strength of the relationship. In text, we also report the range in median values, as well as coefficients of variation in the median ISS per sampled haul or age sample. We report these statistics in order to show how weak and variable these relationships are. We have also added to the discussion section a recommendation that assessment scientists should perform the bootstrap procedure rather than scale either the hauls or nominal sample size to determine age composition ISS.**

Line 416: Change “…the package we built…” to “…the R package we built…”. The DESCRIPTION file in the R package should be updated so that the citation could be filled in so that the information at <https://benwilliams-noaa.github.io/surveyISS/authors.html#citation> is more complete (that same information is provided by the R command citation(“iss”) if the package is installed.

**Response: In the revision we have added ‘R’ before ‘package’ here and are currently working on updating the DESCRIPTION file.**

Line 575: “Number ages” to “Number of ages”

**Response: ‘of’ has been inserted here in the revision.**

Line 580: “plts” to “plots”

**Response: This has been corrected in the revision.**

Line 580: “elipses” to “ellipses”

**Response: As noted in previous responses, we hae taken out the ellipses in the revision and have rather included linear model fits.**

Also, some explanation of how the ellipses were calculated would be useful. Adding a linear regression may also be useful.

**Response: We agree with the reviews suggestion here to add linear regression and have done that in the revision.**

From the editor (Andre Punt) The reviewers raise several very relevant and important points. Reviewer #2 requests consideration of some broader issues. You should either adopt these or provide reasons why this is not appropriate. I am likely to request a 2nd review of the MS by reviewer #2

**Response: As the editor will see in our responses, we have tried to be responsive to all the reviews, particularly to the reviews provided by Reviewer #2. Specifically, we appreciate Reviewer #2’s comment on alternative approaches, as this led to our development of functions to compute the ISS for conditional age-at-length data, which we feel is an important added feature of the R package and the revision. We feel that we have incorporated all the suggestions, with 2 primary exceptions. The first exception being not including results that treat the age samples as random and compute the marginal age compositions directly rather than expand them through the ALK (as pointed out by Reviewer #1). We have provided reasoning above as to why we decided not to include this analysis, as the collections are not completely random and would still need to be expanded through an ALK. We also hypothesize that a similar impact would result when including ageing error whether the age composition is expanded through and ALK or not, and do not think that adding additional analysis and results would really add any value to the revised manuscript. The second exception is providing a table that reports the ISS relationship with the number of hauls or nominal sample size. As we state above, we are reluctant to provide this table because of the lack of a significant relationship between ISS and either hauls or nominal sample size. We would prefer that assessment scientists apply this bootstrap procedure rather than scale the number of hauls or nominal sample size due to the large variability between ISS and these quantities. For either exception, however, if the editor or reviewers strongly feel that the random analysis and hauls-nominal sample size table would be useful, we’d be happy to include them the supplementary material section.**